CLASSIFYING FANTASY VS SCIENCE FICTION POSTS

PROBLEM STATEMENT

- Suppose you love to read fantasy books, and you want to be updated on book reviews and the latest book releases…

- How do we build a model capable of understanding our likes, and then filtering incoming posts to show us what we want?

METHODOLOGY

Datasets

- 1800 posts from science fiction and fantasy subreddits from Jan 27 – 28
- Cross reference datasets with reputable sources
- Fix errors

Data exploration phase & pre-processing

- Remove duplicate posts
- Remove posts with insufficient content
- Removing html, punctuation, stop-words

Modelling

- Naïve Bayes
- Logistic Regression

KEY FINDINGS

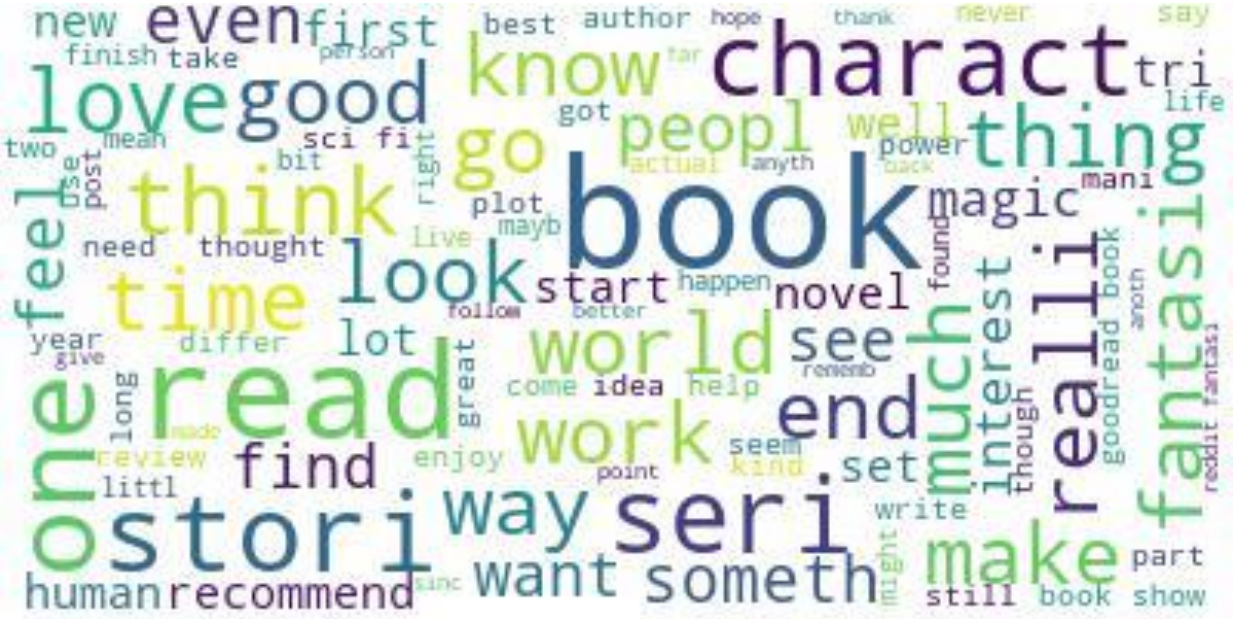
FROM DATA

Naive Bayes model scored

86%

Logistic Regression model scored

85%



Top words from training dataset

CHALLENGES

- Duplicate posts → switched reddits

- Overfitting → Customised stopwords